DOCUMENT RESUME

ED 193 275 TM 800 590

AUTHOR Tyler, Ralph W.
TITLE Educational Evaluation: A Retrospective View.
PUB DATE Sep 80
NOTE 12p.: Paper presented at the Annual Meeting of the American Psychological Association (88th, Montreal, Quebec, Canada, September 1-5, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Educational Assessment; Educational Demand; *Educational History; Educational Innovation; *Educational Objectives; Elementary Secondary Education; Grading; *Program Evaluation; Psychometrics; Public Education; Social Change; *Testing

ABSTRACT

A broad review of educational assessment during the past 40 years (1930's to 1970's) is presented. The review begins with the birth of the term "evaluation", when the depression made educators aware of the need for new programs and ways to appraise them, and continues into an explanation of the Eight-Year Study (1933-41), which gave attention to the assessment of educational objectives and techniques of appraisal. The next milestone discussed is the changed role of public education after World War II. Instruction and the grading system, originally developed to sort students, was modified to encourage as many students as possible to further their education. The final section of the review is a consideration of various checkpoints for program assessment. These points include: (1) assessing what students have learned; (2) comparing ideas of an educational activity with realities to which they refer; (3) checking implementation; (4) monitoring program operations after implementation; and (5) appraising program outcomes according to what the students have learned. (GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PERMISSION TO REPRODUCE THI
MATERIAL HAS BEE.. GRANTED E

R. Tyler

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED193275

Educational Evaluation: A Retrospective View

Ralph W. Tyler

Science Research Associates, Inc.

Director Emeritus, Center for Advanced Study in the Behavioral Sciences

Educational Testing began its phenomenal growth shortly after the end of World War I. It was focused on appraisal of the educational achievement of individual students in relation to other students in the same grade or of the same age. The use of the term evaluation began in educational areas in the 1930's, when the depression brought many students into the upper grades and the high schools and educators became acutely conscious of the need for new programs that would help the new students to learn. But since the programs were previously untried, there was need to evaluate them.

The successful use of psychological and educational tests in the First World War led to their wide adoption by schools and other educational institutions. The Army Alpha test was relatively effective in selecting and classifying the two million young men drafted into military service in 1917-18. After the war, group tests, both of intelligence and achievement, were constructed and developed for school use employing the same methodology that was formulated for the Army Alpha and other military classification tests. The purpose was to sort students, or grade them from excellent to poor. Because the problems that were then recognized were those of developing objective, reliable measures of complex human behavior patterns, primary attention was devoted to methods for constructing so-called objective tests, and techniques for maximizing

TM 800590

discrimination and reliability and for constructing scales of equal units of measure. Standards developed by the testing profession centered on these characteristics. Questions regarding the appropriate purposes for testing, which educational objectives need to be assessed, and which techniques of appraisal and analysis of data are consistent with the purposes, were not given significant attention until the advent of the Eight-Year Study, 1933-41. The following is taken from an article I prepared for the journal, Social Policy - October 1977.

Recognizing Multiple Objectives

The Eight-Year Study stimulated an important change in the conception of what should be assessed in appraising the educational progress of students. This national project involved 30 schools and school systems in an effort to develop educational programs appropriate for the wide range of young people who enrolled in high school during the Great Depression, when unemployment among youth was nearly 100 percent. Many had not planned to go to high school, but there was no place else to go. The high school curriculum had not been designed for this heterogeneous population. These 30 schools were selected as a pilot group to develop new curricula under the condition that they would conduct a careful evaluation of the educational achievement of their students.

At the end of the first year of the study, something of a crisis arose. The high schools were saying that they were being unfairly judged by the standard tests administered to their students. Those traditional tests, they said, were not appropriate to measure the types of learning actually taking place. To develop more appropriate measures,

an evaluation staff was established, and I was asked to serve as director of that staff.

We quickly found that the high school achievement tests then on the market measured only the amount of information which students remembered, or some of the more specific skills used in subjects such as algebra and the foreign languages. The new courses developed in the 30 schools attempted to help students achieve several additional objectives; more effective study skills, more careful ways of thinking, a wider range of significant interests, social rather than selfish attitudes. The available achievement tests did not provide measures of many of the more important outcomes anticipated from these new courses. Furthermore, the content of most significance in the new courses was frequently different from that which had been included before, and the available tests of information did not really measure the new types of information students could be expected to recall from these courses. A comprehensive assessment of the new educational programs could not be carried on unless new means of evaluating student achievement were developed. This was the task undertaken by the evaluation staff.

We found that comprehensive appraisals of the educational programs of schools and colleges were rarely made. Yet, an evaluation of an educational institution is fundamentally the only process by which we find out whether the educational objectives of the institution are being realized. This seems a simple and straighforward task, but there are complexities that were not perceived in 1934. In the first place, since schools generally had not agreed upon their fundamental objectives, there was doubt as to what values schools expected to attain and, therefore, what results to look for in the process of evaluation. Even when

the objectives of a school were agreed upon and stated, they were frequently vague and required clarification in order to be understood. Furthermore, the methods of obtaining evidence about the attainment of some of the objectives were more difficult and less direct than giving a paper-and-pencil test. Finally, the task of summarizing and interpreting the results of an evaluation of a total educational program is complicated. Summaries of educational evaluation were needed for several different groups: for students, teachers, administrators, parents, and patrons. Each of these groups needed somewhat different information, or, at least, it was necessary to present the data in different terms. In spite of the difficulty of these tasks, the evaluation of the Eight-Year Study demonstrated that:

1. Educational institutions could formulate objectives deemed important by faculty, students, parents, administrators, and patrons.
2. These objectives could be clarified by stating them in terms of the patterns of behavior they seek to help students acquire.
3. These patterns of behavior included types that had not previously been appraised by achievement tests: ability to interpret data, to apply principles of science, to use logical principles, social sensitivity, social attitudes, the ability to apply social values and generalizations to social problems, appreciation of literature, appreciation of art, broad and significant interests, and personal-social adjustment.
4. Instruments could be developed that furnish valid evidence of progress toward these varied objectives. The instruments included, in addition to paper-and-pencil tests, observation

guides, interview schedules, questionnaires, inventories, and records made for other purposes. (The Eight-Year Study preceded the marketing of the audio and visual tape equipment and high-speed computers now used in appraising student-achievement.)

5. Human behavior is ordinarily so complex that it was not adequately described or measured by a single term or a simple dimension such as a test score or a grade in a course. Hence, methods of analysis and interpretation of appraisal data were developed to furnish multiple scores, categories, and descriptions to parallel the complexity of the behavior patterns assessed. The study showed that the way in which the student organized his or her behavior patterns was an important aspect to be appraised. There is always a danger that the identification of various types of objectives--a taxonomy--will result in their treatment as isolated behaviors. Thus, the recognition that an educational program seeks to help the student acquire information and develop skills, ways of thinking, attitudes, and interests may result in an assessment program which appraises the development of each of these aspects of behavior separately and makes no effort to relate them. Educators seek to develop in a student a fairly unified reaction pattern: for example, a student who encounters an important socio-civic problem is expected to draw upon her/his store of information, to use already learned skills in locating additional facts, to think through the problem critically, to choose courses of action in terms of fundamental values and attitudes, and to be continually interested in better solutions to such problems.

This clearly involves the relationship of various behavior patterns and their integration. Developing the ability to relate one's various reactions is an important part of a student's education, and a comprehensive assessment should include it.

6. The nature of the assessment influenced the teaching and learning. When students were periodically examined on certain content or skills, a majority of them concentrated their study on this content or these skills even though it was given little or no emphasis in the course of study. Teachers, too, were frequently influenced by their conceptions of what the achievement tests would be appraising. If the tests were thought to emphasize certain points, these points were emphasized in teaching even though they were not included in the plan of the course. This influence of appraisal upon teaching and learning led the evaluation staff to try to develop evaluation instruments in harmony with the new curricula. From the standpoint of influence on teaching and learning, it is better to include crude assessments of progress toward an important objective rather than to make no attempt at its appraisal. An unassessed objective was likely to be one that was given little attention by students and teachers.

This was a reversal of the technical work in educational testing up to that time. The prevailing view was that the refinement of tests of information and subject skills to increase their objectivity, reliability, and precision of measurement was the way to improve educational evaluation.

Oscar K. Buros served as the associate Director of the Evaluation Staff until 1935. He became so much concerned with the lack of accessible information about particular educational and psychological tests that he left the Eight-Year Study to found the Mental Measurements Yearbook. This publication was influential in stimulating revisions in the standards for tests and testing adopted by professional organizations at that time.

Assessment for Education Rather Than Sorting

Another great change in assessment began to take place after World War II. It was brought on by the changed role of public education. In the United States, as long as most jobs were unskilled, and professional and managerial positions comprised a very small sector of the labor force, only a few persons needed to complete secondary school in order to obtain employment and contribute to the economic system. In the political and social spheres the same pyramid prevailed: not more than 5 or 10 percent were expected to be well educated, with the vast majority of the population made up of the "silent electorate." Although this situation had been changing gradually for more than a century, it wasn't until after World War II that rapid changes in the occupational and social structure created sharp new demands for education. For example, in 1800 about 60 percent of the U.S. labor force was engaged in agriculture, whereas in 1900, 38 percent was employed in agriculture--a shift of 22 percent in 100 years. In 1970 agricultural labor was down to about 5 percent--a shift of 33 percent in 70 years. In 1800 the unskilled in all categories comprised more than 80 percent of the labor force, in

1900 they made up about 60 percent and in 1970 about 7 percent. The rapid development of employment in the various services--education, health, recreation, social services, administration, and accounting--has largely taken place since 1948. Now, jobs requiring no schooling are few in number while tasks requiring at least a high school education make up nearly two-thirds of employment opportunities. Schools are now expected to educate all or nearly all children rather than to sort them, using test scores and grades to encourage a fraction of the children to go on with their education and to discourage a large fraction by their lack of marked success in the early years.

The lockstep progress of instruction and the grading system were developed to sort students rather then to help every child and youth get an education. By moving the whole class at the same rate from topic to topic, pacing the movement in terms of the performance of the average students, those with more difficulty in school learning would be certain to get further behind and most of them would give up trying to learn. The resulting feeling of inadequacy was reinforced by the testing and grading system, which year after year gave low marks to those having difficulty. Under the same system those who mastered the school tasks easily and quickly were assigned high marks which encouraged them to go on with their formal education. The testing program was designed to reinforce this sorting procedure. The test results were reported in terms that arranged the students on a continuum from highest to lowest terms, furnishing an apparently objective basis for sorting.

Sorting tests are usually inappropriate for serving education. In constructing a sorting test, it is necessary to find test questions that differentiate among those who take the test. A question that all or

almost all persons answer correctly does not help much to separate the high scoring from the low scoring ones. The same is true for a question that few can answer correctly. Those that help most to discriminate are questions that half of the persons answer correctly and half do not. For this reason, in constructing sorting tests, questions are retained that are at about the fiftieth percentile in difficulty, while most of those that are in the lowest third or the highest third in difficulty are discarded. This produces a test that sorts persons efficiently, but there is no longer a representative sample of the content being tested. Items that most persons have learned are not well represented in the test, nor are items that few persons have learned. Scores on such norm-referenced tests do not show the extent to which students have learned what the schools have to teach, although many lay persons interpret them this way.

Checkpoints for Program Assessment

The changes taking place in the practices as well as the conceptions of assessment are exemplified in the expansion of its definition. This has arisen from the need to obtain and analyze a number of kinds of evidence regarding matters of concern in the planning and operation of an educational institution or program. It includes assessing what students have learned in the institution or program, but is much more than that. In general, it is a means for comparing the ideas involved in an educational activity with realities to which the ideas refer. For example, a program of remedial reading may be planned for the middle grades based on the idea that the pupils have not mastered skills of

word attack and need intensive training in phonics when the real need was for reading materials in these subjects that would offer gradual, sequential practice, with reading selections that would help to build vocabulary and appropriate syntax.

A third stage where a check needs to be made in the development of an educational program is in its implementation. Often schools report that a program is being carried on when, in fact, its basic features are not in operation. Many programs are adopted but not really implemented by schools. Hence, the need to check implementation in evaluating an educational program.

A fourth stage where assessment is helpful is in monitoring the operation of the program after it is implemented. After programs have been in operation for a time, the particular required procedures are sometimes forgotten and the classroom practices revert to those in force before the program was established. This is particularly true for programs that involve procedures quite different from those teachers have previously employed, such as special programs for disadvantaged children or programs emphasizing student self-direction.

A fifth stage in program evaluation is the appraisal of the actual outcomes of the program in terms of what the students have learned.

Educators and Evaluators: Fruitful Interaction

In reviewing the progress of educational assessment during the past 40 years, I am impressed by the dynamic effects of the closer relations between those making new developments in assessment and the particular area in which assessment is needed and will be used. For a generation

after World War I, specialists in testing and assessment, with a few notable exceptions, were segregated as a subdivision of psychology seeking general answers to the problems of educational measurement. In this atmosphere, isolated from educational practice, they developed theories and practices based on older or inadequate conceptions of schooling, of the educational tasks of the school, and of the changing social and educational environment in which learning and teaching are conducted. As a result, the assessment procedures and devices that were published and generally available served only part of the nearly universal needs for assessment that exist throughout the educational world.

As these needs became more pressing, persons whose original training and experience were in various fields of educational practice became involved in evaluation, and, at the same time, persons trained in the specialty of educational evaluation were employed to work in particular areas of education. Out of the interaction stimulated by arrangements of this sort, the theory and practice of educational assessment began to develop more rapidly and relevantly. Upon these foundations we are able to build increasingly effective programs of educational assessment.